Available online at www.sciencedirect.com

ScienceDirect

Transportation Research Procedia 6 (2015) 272 – 284

Transportation Research Procedia

www.elsevier.com/locate/procedia

4th International Symposium of Transport Simulation-ISTS'14, 1-4 June 2014, Corsica, France

Solving large-scale urban transportation problems by combining the use of multiple traffic simulation models

Carolina Osorio*, Krishna Kumar Selvam

Department of Civil and Environmental Engineering, Massachusetts Institute of Technology, Cambridge, MA 02139, USA

Abstract

Transportation agencies often resort to the use of traffic simulation models to evaluate the impacts of changes in network design or network operations. They often have multiple traffic simulation tools that cover the network area where changes are to be made. Nonetheless, these multiple simulators may differ in their modeling assumptions (e.g., macroscopic versus microscopic), in their reliability (e.g., quality of their calibration) as well as in their modeling scale (e.g., city-scale model versus regional-scale model). The choice of which simulation model to rely on, let alone of how to combine their use, is intricate. A larger-scale model may, for instance, capture more accurately the local-global interactions; yet may do so at a greater computational cost. This paper proposes a methodology that enables the simultaneous use of multiple traffic simulation models.

We propose a simulation-based optimization algorithm that embeds information from simulation models with different levels of accuracy and with different levels of computational efficiency. The algorithm combines the use of high-accuracy low-efficiency models with low-accuracy high-efficiency models. This combination leads to an algorithm that can identify points with good performance at a reduced computational cost.

We evaluate the performance of the algorithm with a traffic signal control problem on a small network. We show that the proposed algorithm identifies signal plans with excellent performance, i.e., with reduced average trip travel times, while doing so with a reduction in the computational cost.

Selection and/or peer-review under responsibility of the Organizing Committee of ISTS'14

Keywords: simulation; large-scale optimization; multi-model; queueing theory; signal control

* Corresponding author. Tel.: +1-617-452-3063.
E-mail address: osorioc@mit.edu

Selection and/or peer-review under responsibility of the Organizing Committee of ISTS'14
doi:10.1016/j.trpro.2015.03.021

1. Introduction

Consider a subnetwork within a larger network (e.g., an arterial within a city, a city within a region) where local changes to the supply of the subnetwork are being considered. Transportation agencies often resort to the use of traffic simulators in order to determine the changes to be carried out (e.g., changes in the network design or in the traffic control), and to evaluate the impacts of these changes both locally (i.e., within the subnetwork) as well as globally (i.e. at the larger-scale).

Transportation agencies often have multiple simulators that cover the subnetwork of interest. Nonetheless, these multiple simulators may differ in their modeling assumptions (e.g., one model may be macroscopic, another microscopic), in their reliability (e.g., quality of their calibration) as well as in their modeling scale (e.g., one may be a city-scale model, another a regional model). Most often, transportation experts will consider the advantages and disadvantages of each model, and will ultimately choose one model to rely on in order to determine and study in detail the subnetwork changes. The choice of a model is not an easy task. A larger-scale model may, for instance, capture more accurately the local-global interactions; yet may do so at a greater computational cost.

In this paper, we propose a simulation-based optimization (SO) framework that allows for the combined use of multiple simulation models. We assume that we have access to two models that cover the subnetwork of interest and both have the same modeling assumptions (e.g., same behavioral models). Let R denote the larger scale simulation model (R stands for regional), and C denote the smaller scale simulation model (C stands for city). We assume that the subnetwork of interest where network changes are to be carried out is the full network of C . This subnetwork is entirely modeled in R . This is a scenario which one can easily encounter in practice, where R is an available large-scale model, and C is a smaller model extracted from R , and calibrated based on R outputs. The model R is assumed to lead to more accurate estimates of both local and global performance; yet is significantly more expensive to evaluate.

The family of transportation problems that we consider is continuous and generally constrained problems. The objective function is estimated via a stochastic simulator, whereas the constraints are available in closed-form and are differentiable. Such a problem can be formulated as follows.

$$\min_{x} \quad f(x) = E\left[F(x;\tilde{p})\right] \tag{1}$$

$$\text{subject to} \quad h(x;\tilde{p}) = 0 \tag{2}$$

$$x \in \mathrm{R}^n. \tag{3}$$

In this formulation x represents the decision vector. F is the random variable that describes network performance (e.g., trip travel time). In this formulation the objective function is the expected value of F . The objective function is an unknown function. We can only obtain estimates of it via stochastic simulation. The simulation model is also a function of exogenous parameters (e.g., network topology, calibrated behavioral models) which are represented by $\tilde{p}$. The constraints represented by the function h are available in closed-form. For instance, in the signal control problem considered in this paper they represent green time constraints for every intersection (e.g., bounds, linear constraints).

In this paper, we assume that the objective function accounts for the local (i.e., subnetwork) performance. That is, the aim is to improve local conditions. We propose a simulation-based optimization (SO) technique that allows for simultaneous use of both simulators, R and C , in order to address such a problem. The goal is to achieve a suitable trade-off between obtaining accurate local performance estimates and the associated computational costs.

Recent reviews of traffic simulation models include Barceló (2010), Ratrout and Rahman (2009). The main families of models are known as macroscopic, mesoscopic and microscopic. Microscopic simulation (also referred to as microsimulation) models are the most detailed, yet also the most computationally inefficient models. Their computational inefficiency limits their use to address large-scale transportation problems. On the other hand, macroscopic models are computationally more efficient and hence are often used for the analysis of large-scale

problems. Nonetheless, they do not embed a detailed representation of travel demand, network supply, and traffic dynamics.

In order to achieve a suitable trade-off between modeling detail and computational efficiency, the transportation community has mostly resorted to the formulation of mesoscopic models, which combine ideas from macroscopic and microscopic models. Nonetheless, this leads to models that are neither as detailed as microscopic models nor as efficient as macroscopic models. A second approach has been to develop frameworks that combine the use of multiple traffic simulation models. For a detailed literature review on these topics, see Osorio and Selvam (2014).

Hybrid traffic models can be classified broadly into two categories. The first category models different areas of the network with different modeling scales (e.g., microscopic or macroscopic). Work in this field includes: Sewall et al. (2011), Burghout (2004), Horowitz (2004), Bourrel and Lesort (2003), Magne et al. (2000). Much of the past work in this category focuses on the issue of model consistency at the multi-scale boundaries, and in particular, on the topic of aggregation and disaggregation (e.g., to achieve consistency between vehicle flows and densities at the macroscopic scale and individual vehicles at the microscopic scale).

The second category allows for areas of the network to be simultaneously modeled with multiple scales. Work in this area includes, for instance, Rousseau et al. (2008), Bunch et al. (1999), Montero et al. (1998), Van Vliet and Hall (1997). Typically, a large region is modeled with low fidelity (e.g., macroscopic approach), and a smaller subnetwork within the larger region is separately modeled with higher fidelity (e.g., microscopic/mesoscopic) model. This means that the smaller subnetwork is modeled with both the low and the high fidelity models. The boundary conditions of the smaller scale model (e.g., origin-destination matrix) are typically estimated based on large-scale low-fidelity outputs.

This paper considers this second category. Nonetheless, it uses two high-fidelity microscopic traffic models. That is, both the larger region and the smaller subnetwork are modeled with the same high-fidelity resolution. Nonetheless, the larger region model is computationally costly to evaluate, whereas the subnetwork model is more efficient but less accurate since it considers fixed boundary conditions. Hence, this paper proposes a framework that combines the use of both models in order to address transportation optimization problems in a computationally efficient way.

Multi-model optimization frameworks have been proposed in various fields other than transportation (e.g., Alexandrov et al. 1999). Nonetheless, past work has considered either the use of several analytical models with varying computational costs, or the use of several deterministic simulation-based models with various costs. This work integrates ideas from such frameworks within a context where the models consist of various stochastic simulators.

2. Methodology

2.1. General framework

Our past work in the field of simulation-based optimization (SO) has combined the use of an analytical traffic model with a stochastic microscopic traffic simulator (Osorio and Bierlaire 2013). This general idea has been used to address a variety of simulation-based signal control problems including large-scale problems (Osorio and Chong 2014), emissions and fuel-efficient problems (Osorio and Nanduri 2014b, a), as well as travel time reliability problems (Chen et al. 2013).

Recall from Section 1 that we consider multiple simulation models: a larger scale model R and a smaller subnetwork model C. The aim of this paper is to derive a transportation strategy (e.g., a signal control plan, a network design alternative; hereafter called a point) that provides subnetwork improvement when evaluated with R. We assume we have a fixed simulation run-time budget. The objective is to identify a point with improved performance within that budget.

There are three types of techniques:

1. Use only R
2. Use only C
3. Use a combination of R and C.

The first strategy will definitely lead to a strategy with improved performance, yet is inefficient since R takes longer to execute. This first strategy will therefore evaluate the performance of fewer points within the computational budget. The second strategy enables the largest number of simulation runs within the budget, yet may not lead to a strategy with improved performance when evaluated with R. This is because the boundary conditions of C are fixed, and do not vary across points. Hence, these boundary conditions may be inaccurate, leading to inaccurate performance measure estimates. The third strategy is that proposed in this paper. It attempts to reach a trade-off between the two other strategies.

The methodology is defined for a general transportation optimization problem. In this paper, we illustrate its formulation and use with a traffic signal control problem. We use interchangeably the terms signal plan and point.

We assume we have access to a calibrated large-scale model R. We calibrate the subnetwork model C based on the outputs of R and do so for a given signal plan (e.g., calibration of behavioral parameters, of origin-destination (OD) matrix). This is done once, before starting the optimization algorithm.

A one-shot calibration of C before running the optimization algorithm means that the exogenous parameters of C (i.e., its boundary conditions) are fixed. Hence, the accuracy of the performance estimates provided by C will vary depending on the signal plans that are evaluated. For example, changes in the signal plans within the subnetwork may lead to changes in the OD matrix of C. Since these changes are not accounted for by C, it can lead to estimates that are less accurate than those of R. Extensions of this framework may calibrate C iteratively throughout the optimization, such as to have point-specific (i.e., signal plan-specific) boundary conditions.

At each iteration of the SO algorithm, the main decision to be made is which simulator to call (R or C). Given the varying accuracy of C, we propose an approach that approximates the accuracy loss of C. In other words, we analytically approximate the changes in the exogenous parameters of C, as well as their effect on the objective function. We do this by using an analytical traffic model that covers the full R network.

At iteration k of the SO algorithm, let x_k denote the signal plan that is to be simulated. Let θ_0 denote the exogenous (i.e., fixed) parameters of C (in this paper they are fixed throughout the entire optimization process). For signal plan x_k, let $\hat{\theta}_k$ denote the approximated value of the parameters of C (e.g., OD matrix of C that results from the signal plan x_k). In this paper, the approximation of $\hat{\theta}_k$ is derived by the use of an analytical traffic model. Let $g\left(x;\hat{\theta}_k\right)$ denote the approximation of the objective function provided by the analytical traffic model for signal plan x and parameters $\hat{\theta}_k$. In other words, g approximates the function f of Equation (1).

We use the following test to determine whether at iteration k we can rely on the estimates provided by C. If

$$\left|g\left(x;\hat{\theta}_k\right)-g\left(x;\theta_0\right)\right|<\eta, \tag{4}$$

then run C; otherwise run R.

That is, if the error in prediction as approximated by the analytical model under the two parameter values is below a threshold, then we expect C to provide an accurate performance estimate; and since it can do so at cheaper cost, then it is the best approach. Otherwise, we run the more computationally expensive yet more accurate R model. The algorithm stops once the computational budget has been reached.

The general algorithm used in this paper is a metamodel SO algorithm. We extend the algorithm of Osorio and Bierlaire (2013) in order to allow for the use of multiple traffic simulators. The main idea is presented in Fig. 1. The algorithm is based on the derivative-free trust region algorithm of Conn et al. (2009). At each iteration of the algorithm simulation observations are collected. They are used to fit a metamodel (denoted m). The metamodel is used to derive a point (e.g., a signal plan), often referred to as a trial point, that is expected to provide improved performance. The choice of which simulator to run is carried out, and the selected simulator is used to estimate the performance of the trial point. In Fig. 1, the simulation-based estimate is denoted $\hat{f}\left(x\right)$. These new simulation observations are used to re-fit the metamodel, and improve its accuracy. These steps are repeated iteratively.

The algorithm used in this paper differs from that of Osorio and Bierlaire (2013) in two ways. First, in Osorio and Bierlaire (2013) a single simulator is used. Here we have included the simulator selection step. Second, in this paper we use a quadratic polynomial as the metamodel, whereas Osorio and Bierlaire (2013) uses a more complex

metamodel formulation. For all algorithmic details, we refer the reader to Osorio and Selvam (2014) and to Osorio and Bierlaire (2013).

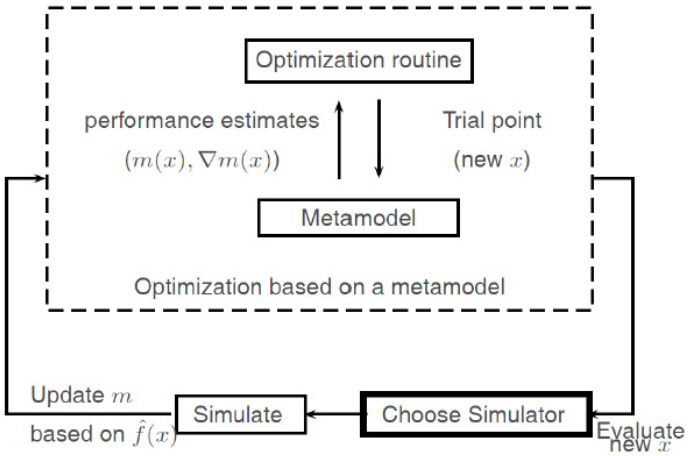

Fig. 1. SO framework that allows for multiple simulation models. Adapted from Alexandrov et al. (1999).

2.2. Analytical error approximation

In this section, we present the formulation of the analytical model g (of Equation 4). The formulation builds upon the traffic model of Osorio (2010, Chap. 4). First, we present the initial model of Osorio (2010), then we present its extension.

In Osorio (2010), a road network is modeled as a finite (space) capacity queueing network. Each lane is modeled as one (or a set) of queues. Each queue is considered an $M/M/1/\ell$ queue, where ℓ is the space capacity of the queue. This space capacity represents an upper bound on the queue-length, and is used to capture the propagation of congestion (e.g., queue spillbacks).

For a given queue i, the following notation is used.

γ_i	external arrival rate;
λ_i	total arrival rate;
μ_i	service rate;
$\tilde{\mu}_i$	unblocking rate;
$\hat{\mu}_i$	effective service rate;
ρ_i	traffic intensity;
P_i^f	probability of being blocked after service at queue i;
ℓ_i	space capacity;
N_i	number of vehicles in queue i;
$P(N_i = \ell_i)$	probability of queue i being full;
p_{ij}	transition probability from queue i to queue j;
D_i	set of downstream queues from queue i

For a given road network represented as a queueing network, the marginal queue-length distributions of each queue are obtained by simultaneously solving for all queues the following system of equations.

$$\lambda_i = \gamma_i + \frac{\sum_j p_{ji}\lambda_j P\left(N_j < \ell_j\right)}{P\left(N_j < \ell_i\right)} \tag{5}$$

$$\frac{1}{\hat{\mu}_i} = \frac{1}{\mu_i} + P_i^f \frac{1}{\tilde{\mu}_i} \tag{6}$$

$$\frac{1}{\tilde{\mu}_i} = \sum_{j \in \mathrm{D}_i} \frac{\lambda_j P\left(N_j < \ell_j\right)}{\lambda_i P\left(N_i < \ell_i\right)\hat{\mu}_j} \tag{7}$$

$$P\left(N_i = \ell_i\right) = \frac{1-\rho_i}{1-\rho_i^{\ell_i+1}}\rho_i^{\ell_i} \tag{8}$$

$$P_i^f = \sum_j p_{ij} P\left(N_j = \ell_j\right) \tag{9}$$

$$\rho_i = \frac{\lambda_i}{\hat{\mu}_i} \tag{10}$$

We briefly describe the interpretation of these equations. Equation (5) describes the conservation of flow between upstream and downstream queues. For queue i, its total arrival rate, λ_i, is related to its external arrival rate, γ_i and to the arrivals arising from upstream queues (second-term in the right-hand side of the equation). Equation (6) describes the service process of a vehicle, which is composed of two phases. First, the vehicle undergoes an initial service. The queue has an underlying service rate, μ_i, that is determined by its underlying supply (e.g., flow capacity of the downstream intersection). After receiving service, a vehicle that is at queue i and is ready to proceed to queue j may do so if queue j is not full. If queue j is full (i.e., if there is a spillback at queue j), then the vehicle is forced to remain at queue i. This is known in queueing theory as blocking. This occurs with probability P_i^f and this second service is referred to as blocking time, the expected blocked time is given by $1/\tilde{\mu}_i$. Equation (7) describes the expected blocking time, which is a function of the effective service rate of downstream queue j, $\hat{\mu}_j$, (i.e., its spillback dissipation rate). Equation (8) describes the probability that queue i is full, it is also known as the blocking probability. In vehicular traffic this represents the spillback probability. The expression of Equation (8) is obtained by assuming that queue i is an $M/M/1/\ell$ queue (e.g., Bocharov et al. 2004). Equation (9) describes the probability that a vehicle at queue i gets blocked (i.e., that it cannot proceed downstream of queue i due to downstream spillbacks). Equation (10) describes the traffic intensity, which is a ratio of demand to supply.

The main limitation of the model of Osorio (2010) for the purpose of our work is that it assumes exogenous turning probabilities, p_{ij}. In this paper, the purpose of the analytical model is to approximate how subnetwork boundary conditions may change due to changes in supply. More specifically, we want to approximate how the OD matrix of the subnetwork changes due to changes in the subnetwork signal plans. Hence, accounting for endogenous assignment is necessary.

In this paper, the turning probabilities, p_{ij} are considered endogenous and are derived by considering a multinomial logit route choice model. The analytical formulation of the assignment is formulated as follows.

d_s demand for OD pair s;
c_t travel cost of path t;
y_t flow on path t;
r_{ti} proportion of flow on path t that goes through queue i;
a_{ti} indicates whether path t contains queue i;
a^*_{ti} indicates whether the first link of path t is link i;
z_{ij} indicates whether queue i and queue j are parallel queues within the same link;
l_{st} probability that a vehicle travelling the OD pair s takes path t;
$E[T_i]$ travel cost of queue i;
$E[N_i]$ number of vehicles in queue i;
l^{veh} average vehicle length;
$v^{freeflow}$ free flow speed;
θ route choice probability parameter;
P_s set of paths of OD pair s;
S set of OD pairs;
Q set of queue indices;
T set of paths indices;
G_{ij} set of paths that consecutively go through queues i and j;
H_i set of paths that go through queue i;

$$p_{ij} = \frac{\sum_{t \in \mathrm{G}_{ij}} y_t}{\sum_{t \in \mathrm{H}_i} y_t} \qquad \forall i \in \mathrm{Q}, \forall j \in \mathrm{Q} \tag{11}$$

$$y_t = \sum_{s \in \mathrm{S}} d_s l_{st} \qquad \forall t \in \mathrm{T} \tag{12}$$

$$l_{st} = \frac{e^{-\theta c_t}}{\sum_{j \in \mathrm{P}_s} e^{-\theta c_j}} \qquad \forall s \in \mathrm{S}, \forall t \in \mathrm{P}_s \tag{13}$$

$$c_t = \sum_{i \in \mathrm{Q}} r_{ti} E[T_i] \qquad \forall t \in \mathrm{T} \tag{14}$$

$$r_{ti} = \frac{a_{ti}}{\sum_{j \in \mathrm{Q}} a_{tj} z_{ij}} \qquad \forall t \in \mathrm{T}, \forall i \in \mathrm{Q} \tag{15}$$

$$E\left[T_i\right]=\frac{E\left[N_i\right]}{\lambda_i P\left(N_i<\ell_i\right)}+\frac{l^{veh}\left(\ell_i-E\left[N_i\right]\right)}{v^{freeflow}} \qquad \forall i\in \mathrm{Q} \tag{16}$$

$$E\left[N_i\right]=\frac{\rho_i}{1-\rho_i}-\frac{\left(\ell_i+1\right)\rho_i^{\ell_i+1}}{1-\rho_i^{\ell_i+1}} \qquad \forall i\in \mathrm{Q} \tag{17}$$

$$\gamma_i=\sum_{t\in \mathrm{T}} a_{ti}^{*} r_{ti} y_t \qquad \forall i\in \mathrm{Q} \tag{18}$$

The indicator variables are defined as follows.

$$a_{ti}=\begin{cases}1 & \text{if queue } i \text{ is part of path } t,\\ 0 & \text{otherwise.}\end{cases} \tag{19}$$

$$a_{ti}^{*}=\begin{cases}1 & \text{if queue } i \text{ is part of the first link (road) of path } t,\\ 0 & \text{otherwise.}\end{cases} \tag{20}$$

$$Z_{ij}=\begin{cases}1 & \text{if queue } i \text{ and queue } j \text{ belong to the same link, i.e., parallel lanes;}\\ 0 & \text{otherwise.}\end{cases} \tag{21}$$

Equation (11) defines the probability of turning from queue i to queue j as the ratio of the total flow along paths that have queues i and j as consecutive queues and of the total flow that goes through queue i. Equation (12) defines the flow along a path t, it is a function of the total demand of a given OD-pair s, denoted d_s, and the probability of choosing path t for OD-pair s, denoted l_{st}. Note that the OD-pair demand d_s of the full R network is exogenous, and obtained from the OD matrix of the R model. The path choice probability is given by the multinomial logit expression of Equation (13). The deterministic component of the utility function for a given route t is defined as a function of a single (exogenous) parameter θ and a single route cost c_t. The latter is defined by Equation (14) as the expected travel time for route t. The route travel time is a function of queue travel time $E\left[T_i\right]$, which is given by Equation (16). In Equation (16), the first term on the right-hand side represents the (expected) delay at queue i, it is obtained by applying Little's law (Little 2011, 1961) to a finite (space) capacity queue. The second term is an approximation of the travel time to reach the physical queue of vehicles: the numerator approximates the available road-space length not occupied by a stationary vehicular queue; the denominator is the roads free-flow speed. Equation (17) represents the expected number of vehicles in queue i, $E\left[N_i\right]$, the derivation of this closed-form expression is given in Appendix A of Osorio and Chong (2014). Equation (18) gives the expression for the external arrival rate of queue i, γ_i. In the model of Osorio (2010), this rate is exogenous. In this paper, since we account for endogenous assignment, the external arrival rates of a given queue depend on the (endogenous) path choice probabilities, and hence are themselves endogenous.

The System of Equations (5)-(18) describes the analytical model with endogenous assignment. It is used to approximate the expression in Equation (4), which represents the accuracy loss due to approximating the objective function with the subnetwork simulation model while not accounting for the change in the boundary conditions of the subnetwork. In other words, the function g of Equation (4) is obtained by evaluating the analytical traffic model (Equations (5)-(18)) and using it to approximate the (simulation-based) objective function f (of Equation (1)). In the next section, we detail how this is done with a specific example of a signal control problem.

3. Case study

3.1. Signal control problem

In this section, we consider a fixed-time traffic signal control problem. To formulate the problem, we introduce the following notation.

b_i	available cycle ratio of intersection i;
x_L	vector of minimum green splits for each phase;
I	set of intersection indices;
$\mathrm{PH}(i)$	set of phase indices of intersection i;
$T_{sub}(x)$	subnetwork travel time for the signal plan x;

$$\min_{x} \quad f(x) = E\left[T_{sub}(x)\right] \tag{22}$$

$$\text{subject to} \quad \sum_{j \in \mathrm{PH}(i)} x(j) = b_i, \qquad \forall i \in I \tag{23}$$

$$x \geq x_L. \tag{24}$$

The objective function of this problem is the expected travel time in the subnetwork, T_{sub} represents the subnetwork travel time, and x is the decision vector of green splits for all endogenous signal phases. The parameter b_i is an exogenous parameter that represents the proportion of cycle time that can be allocated. This proportion excludes any fixed times (e.g., all-red times). Equation (23) ensures that for each intersection all available green time is allocated across all phases. In this equation $x(j)$ is the j^{th} element of x, it represents the green split of signal phase j. Lower bounds for the green splits are ensured through (24).

3.2. Analytical approximation of the accuracy loss

At a given iteration k of the SO algorithm, the simulators are used to estimate the performance of a signal plan x_k. In other words, the simulators are used to esimate $f(x_k)$ (as defined in Equation (22)). We use the analytical traffic model presented in Section 2 to approximate this objective function in two ways.

First, we approximate its value assuming the boundary conditions of the subnetwork are unchanged, this is denoted by $g(x;\hat{\theta}_0)$ in Equation (4). In particular, the subnetwork OD-demand and the subnetwork assignment are equal to their initial values $\hat{\theta}_0$. In the analytical traffic model this means that the external arrival rates γ_i take their initial values, and are exogenous. Assuming these exogenous values, we can solve the System of Equations (5)-(10), and then evaluate the corresponding analytical objective function with Equations (16)-(17). Let $f_C(x_k)$ denote this analytical approximation of $f(x_k)$. Then:

$$f_C(x_k) = \sum_{i \in \mathrm{A}} E\left[T_i\right], \tag{25}$$

where A denotes the set of queues in the subnetwork. We denote this approximation f_C since it approximates the estimate of the simulator C, which assumes a fixed subnetwork OD matrix.

Second, we analytically approximate $f(x_k)$ accounting for the fact that the subnetwork boundary conditions may change, this is denoted by $g(x;\hat{\theta}_k)$ in Equation (4). We use the analytical traffic model with endogenous

subnetwork external arrival rates γ_i, we solve the System of Equations (5)-(18). Let $f_R(x_k)$ denote this analytical approximation of $f(x_k)$. Then:

$$f_R(x_k) = \sum_{i \in A} E[T_i]. \tag{26}$$

We denote this approximation f_R since it approximates the estimate of the simulator R, where subnetwork travel demand indeed varies with the subnetwork signal plans.

Equation (4) is then obtained by evaluating:

$$\left| f_R(x_k) - f_C(x_k) \right| < \eta. \tag{27}$$

If this inequality holds, then we assume that the error made by the simulator C is small, hence we choose to run simulator C. Otherwise, we prefer to run the more accurate yet more costly-to-run simulator R.

3.3. Example

We illustrate the performance of the proposed algorithm with a small toy network example. Fig. 2 displays the networks of interest. The top network represents the full network R, which considers two OD pairs: $A \to B$ and $C \to D$. Each OD pair of network R has two path alternatives, one of which goes through a signalized intersection (denoted by the yellow square). The subnetwork C is displayed to the bottom of Fig. 2. It contains the same OD pairs as the R network, but accounts only for the paths that travel through the intersection, i.e., it considers two OD pairs, each with one path. A change in the signal plan of the intersection may affect the path choice probabilities. This change will be reflected when running simulator R but will not be reflected when running simulator C. The three strategies described in Section 2.1 were tried out with these two models.

Recall that the objective function is the expected subnetwork travel time, i.e., the travel time in the links of subnetwork C. The intersection has two endogenous signal phases: one for east-west bound traffic and the second for north-south bound traffic. We evaluate the performance of the three strategies described in Section 2.1. We allow for a maximum of 21 simulation runs. That is approach 1 (resp. 2) allows for 21 runs of simulator R (resp. C), while approach 3 allows for a total of 21 runs, which consist of a combination of runs from R and from C.

We consider three different initial signal plans. For each initial signal plan and each approach, we run the SO algorithm 3 times, allowing each time for a maximum of 21 simulation runs. Each time we run the SO algorithm, we obtain a new signal plan proposed by the SO algorithm. In order to evaluate the performance of this proposed signal plan, we embed it within the R simulator and run 50 simulation replications. We then plot the empirical cumulative distribution function (cdf) of these 50 simulation replications. Fig. 3 displays the corresponding cdf curves. Each plot of the figure considers a different initial signal plan. Each plot contains 8 cdf curves:

- The black curve corresponds to the initial signal plan.
- The 3 blue curves correspond to the 3 signal plans proposed by only running the C simulator. This is the least accurate yet also the least computationally costly approach.
- The 3 red curves corresponding to the 3 signal plans proposed by only running the R simulator. This is the most accurate yet also the most computationally costly approach.
- The 3 green curves correspond to the 3 signal plans proposed by our approach.

For each plot, the x-axis represents the average subnetwork travel time. For a given x value the corresponding y value of the curve represents the proportion of replications (out of the 50 replications) where the simulated average travel time was smaller than x. The more the cdf curves are shifted to the left, the higher the proportion of simulated observations with low average subnetwork travel times.

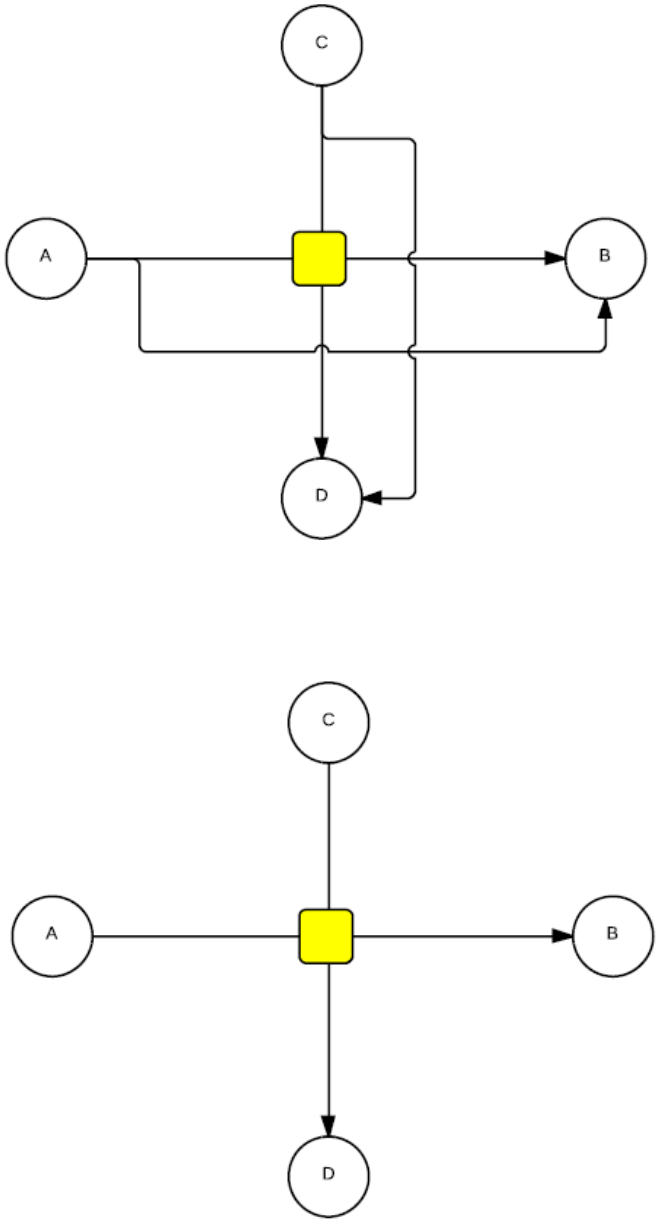

Fig. 2. Network topologies: for the R network (top network) and C network (bottom network).

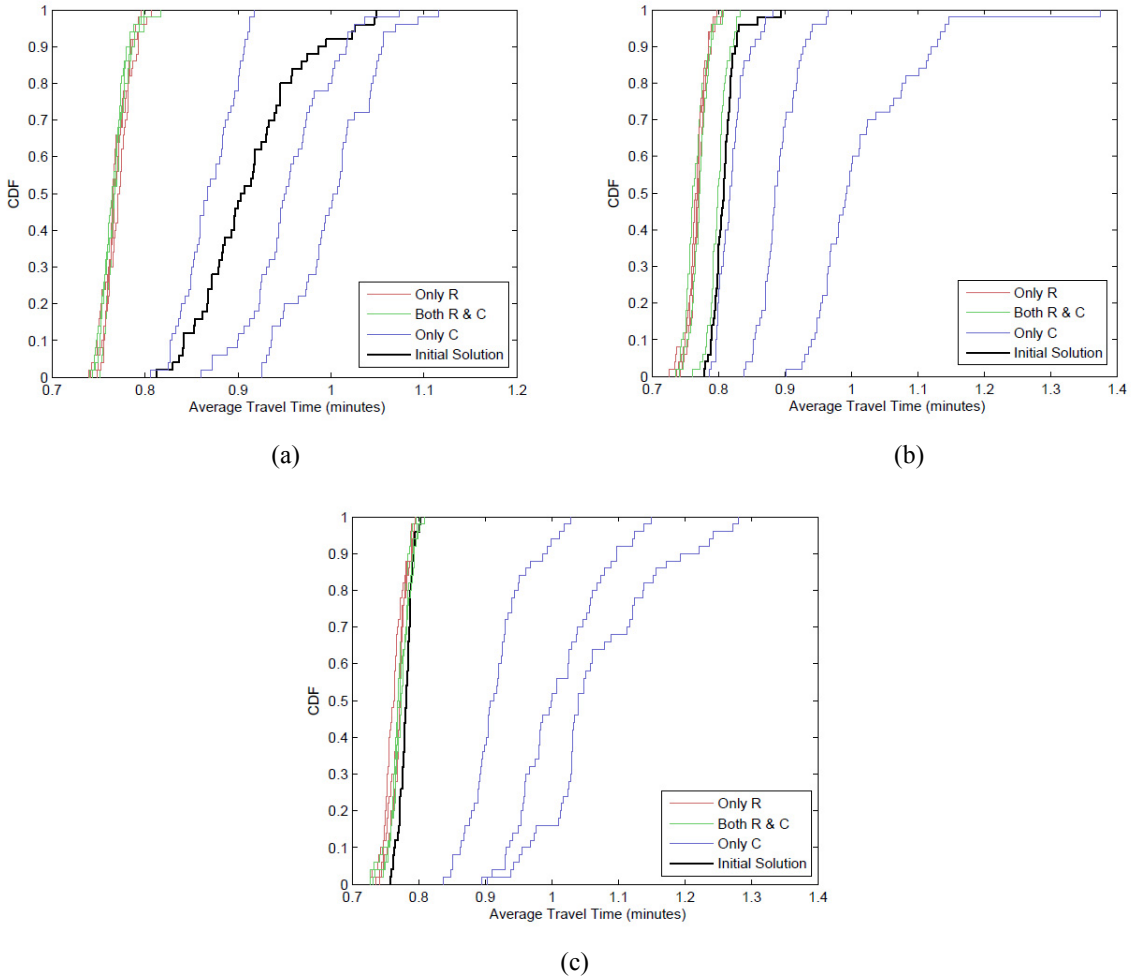

Fig. 3. Empirical cumulative distribution functions of the average subnetwork travel time, for each signal plan proposed by each of the three approaches.

When running only the simulator C, all 3 plots of Fig. 3 indicate that signal plans with poor performance are derived. In particular, in Fig. 3(a) two out of the 3 plans proposed by C perform worse than the initial signal plan, in Fig. 3(b) and Fig. 3(c) all 3 proposed plans are worse than the initial plan. In Fig. 3(c) all 3 proposed plans perform significantly worse than the initial plan.

When running only the simulator R, all 3 plots of Fig. 3 indicate that signal plans with good performance are obtained (with a subnetwork travel time average of approximately 0.75 minutes). Note that for Fig. 3(c) the improvement of the signal plans proposed by R compared to the initial plan is not large, yet the initial signal plan already had a good performance with low average travel times.

When running a combination of simulators R and C, the signal plans systematically yield performance similar to the signal plans propose by running only R. Additionally, for the proposed approach the R model was called on average 66% of the time, while the C model was called 34% of the time.

These results indicate that the proposed approach identifies signal plans with good performance and does so at a lower computational cost. Ongoing results, to be presented at the conference, consider a large-scale network where the run-time of R (resp. C) is in the order of 90 seconds (resp. 2.8 seconds). The proposed approach yields signal

plans with performance just as good as those obtained by only running R, yet it only runs R an average of 42% of the time. This represents a 56% reduction in run-time for a given run of the SO algorithm.

4. Conclusion

This paper presents a simulation-based optimization methodology that enables the combined use of multiple stochastic simulators. This combination allows to trade-off the high computational costs of running accurate large-scale simulators with the lower costs of running less accurate smaller-scale simulators. We illustrate the proposed approach with a signal control problem on a small network example. The proposed approach identifies signal plans with good performance and can do so at a lower computational cost than when systematically running the larger-scale simulator. Ongoing results, to be presented at the conference, consider a large-scale network. We show that we can identify signal plans with good performance while reducing the computational run-time by 56%.

References

Alexandrov, N. M., Lewis, R. M., Gumbert, C. R., Green, L. L., and Newman, P. A., 1999. Optimization with variable-fidelity models applied to wing design. Technical Report CR-1999-209826, NASA Langley Research Center, Hampoton, VA, USA.

Barceló, J., 2010. Fundamentals of traffic simulation, volume 145 of International Series in Operations Research and Management Science. Springer, New York, USA.

Bocharov, P. P., D'Apice, C., Pechinkin, A. V., and Salerno, S., 2004. Queueing theory, chapter 3, pages 96-98. Modern Probability and Statistics. Brill Academic Publishers, Zeist, The Netherlands.

Bourrel, E. and Lesort, J. B., 2003. Mixing microscopic and macroscopic representations of traffic flow: Hybrid model based on lighthill-whitham-richards theory. Transportation Research Record: Journal of the Transportation Research Board, 1852(1), 193-200.

Bunch, J. A., Hatcher, S. G., Larkin, J., Nelson, G. G., Proper, A. T., Roberts, D. L., Shah, V., and Wunderlich, K. E., 1999. Incorporating its into corridor planning: Seattle case study. Technical report.

Burghout, W., 2004. Hybrid microscopic-mesoscopic traffic simulation.

Chen, X., Osorio, C., and Santos, B., 2013. Travel time reliability in signal control problem: Simulation-based optimization approach. In proceedings of the Transportation Research Board (TRB) Conference, Washington DC, USA.

Conn, A. R., Scheinberg, K., and Vicente, L. N., 2009. Global convergence of general derivative-free trust-region algorithms to first- and second-order critical points. SIAM Journal on Optimization, 20(1), 387-415.

Horowitz, R., 2004. Development of integrated meso/microscale traffic simulation software for testing fault detection and handling in ahs.

Little, J. D. C., 1961. A proof for the queueing formula: . Operations Research, 9(3), 383-387.

Little, J. D. C., 2011. Little's law as viewed on its 50th anniversary. Operations Research, 59(3), 536-549.

Magne, L., Rabut, S., and Gabard, J. F., 2000. Towards an hybrid macro-micro traffic flow simulation model. In INFORMS Salt Lake City Spring 2000 Conference.

Montero, L., Codina, E., Barceló, J., and Barceló, P., 1998. Combining macroscopic and microscopic approaches for transportation planning and design of road networks. In Proceedings of the 19th ARRB Transport Research Conference, Sydney.

Osorio, C., 2010. Mitigating network congestion: analytical models, optimization methods and their applications. Ph.D. thesis, École Polytechnique Fédérale de Lausanne.

Osorio, C. and Bierlaire, M., 2013. A simulation-based optimization framework for urban transportation problems. Operations Research, 61(6), 1333-1345.

Osorio, C. and Chong, L., 2014. A computationally efficient simulation-based optimization algorithm for large-scale urban transportation. Transportation Science. Forthcoming.

Osorio, C. and Nanduri, K., 2014a. Emissions mitigation: coupling microscopic emissions and urban traffic models for signal control. Submitted.

Osorio, C. and Nanduri, K., 2014b. Energy-efficient urban traffic management: a microscopic simulation-based approach. Transportation Science. Forthcoming.

Osorio, C. and Selvam, K. K., 2014. Multi-model simulation-based optimization. Technical report, Department of Civil and Environmental Engineering, Massachusetts Institute of Technology (MIT).

Ratrout, N. T. and Rahman, S. M., 2009. A comparative analysis of currently used microscopic and macroscopic traffic simulation software. Arabian Journal for Science & Engineering (Springer Science & Business Media BV), 34.

Rousseau, G., Scherr, W., Yuan, F., and Xiong, C., 2008. An implementation framework for integrating regional planning model with microscopic traffic simulation. In Logistics: the emerging frontiers of transportation and development in China: Proc. 8th Int. Conf. Chinese Logistics and Transportation Professionals.

Sewall, J., Wilkie, D., and Lin, M. C., 2011. Interactive hybrid simulation of large-scale traffic. In ACM Transactions on Graphics (TOG), volume 30, page 135. ACM.

Van Vliet, D. and Hall, M., 1997. Saturn 9.3-user manual. The Institute for Transport Studies, University of Leeds, Leeds, 3.